Introduction to Machine Learning

Introduction to Linear Classifiers

Andres Mendez-Vazquez

February 11, 2019

Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $\boldsymbol{w}^T\boldsymbol{x} + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Outline

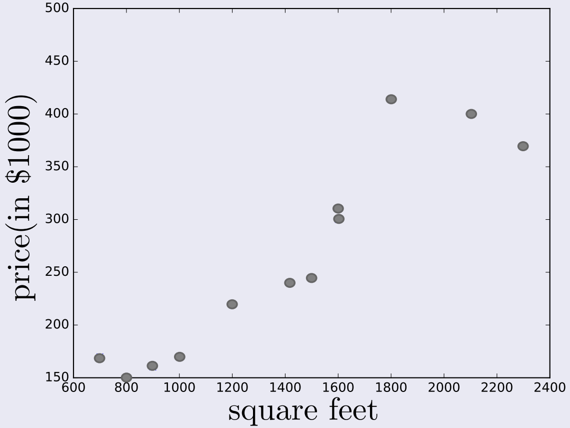
1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Many Times

We have this kind of data sets (House Prices)

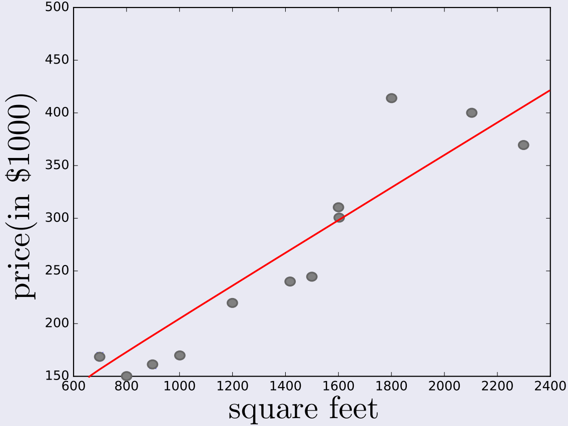
$$\begin{pmatrix} \text{Squared Feet} \\ \text{Price} \end{pmatrix} \rightarrow \begin{pmatrix} 2104 \\ 400 \end{pmatrix} \begin{pmatrix} 1800 \\ 460 \end{pmatrix} \begin{pmatrix} 1600 \\ 300 \end{pmatrix} \begin{pmatrix} 2300 \\ 370 \end{pmatrix} \cdots$$



Thus

We can adjust a line/hyperplane to be able to forecast prices

$$\begin{pmatrix} \text{Squared Feet} \\ \text{Price} \end{pmatrix} \rightarrow \begin{pmatrix} 2104 \\ 400 \end{pmatrix} \begin{pmatrix} 1800 \\ 460 \end{pmatrix} \begin{pmatrix} 1600 \\ 300 \end{pmatrix} \begin{pmatrix} 2300 \\ 370 \end{pmatrix} \cdots$$



Thus, Our Objective

To find such hyperplane

To do forecasting on the prices of a house given its surface!!!

Here, where "Learning" Machine Learning style comes around

Basically, the process defined in Machine Learning!!!

Thus, Our Objective

To find such hyperplane

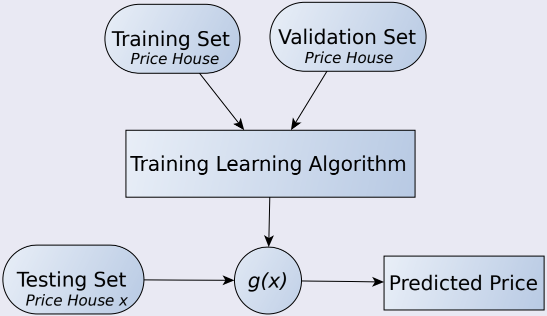
To do forecasting on the prices of a house given its surface!!!

Here, where "Learning" Machine Learning style comes around

Basically, the process defined in Machine Learning!!!

Then, in Supervised Training

We have the following process



Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector
2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

What is it?

First than anything, we have a parametric model!!!

Here, we have an hyperplane as a model:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \tag{1}$$

Note: $\boldsymbol{w}^T\boldsymbol{x}$ is also know as dot product

In the case of $\mathbb{R}^2$

We have:

$$g(x) = (w_1, w_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + w_0 = w_1x_1 + w_2x_2 + w_0 \tag{2}$$

What is it?

First than anything, we have a parametric model!!!

Here, we have an hyperplane as a model:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 \quad (1)$$

Note: $\boldsymbol{w}^T \boldsymbol{x}$ is also know as dot product

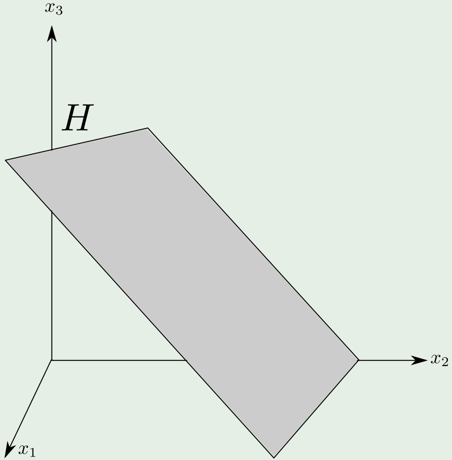
In the case of $\mathbb{R}^2$

We have:

$$g(\boldsymbol{x}) = (w_1, w_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + w_0 = w_1 x_1 + w_2 x_2 + w_0 \quad (2)$$

Example

Hyperplane in $\mathbb{R}^3$



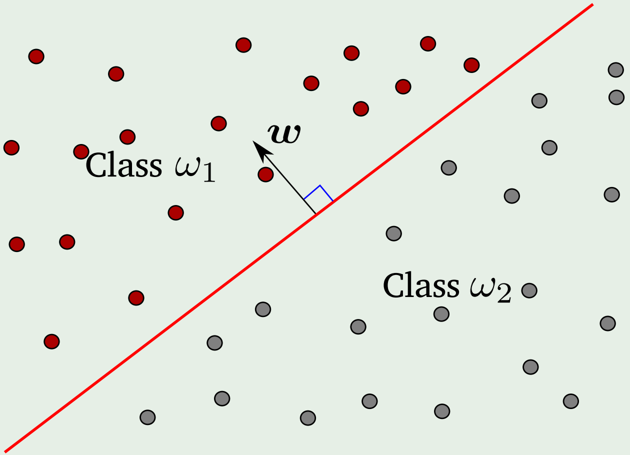
Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Splitting The Space $\mathbb{R}^2$

Using a simple straight line (Hyperplane)



Splitting the Space?

For example, assume the following vector $\boldsymbol{w}$ and constant w_0

$$\boldsymbol{w} = (-1, 2)^T \text{ and } w_0 = 0$$

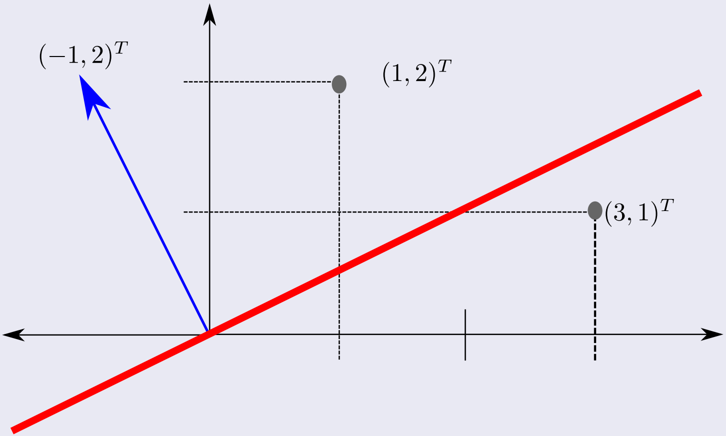
Hyperplane

Splitting the Space?

For example, assume the following vector $\boldsymbol{w}$ and constant w_0

$$\boldsymbol{w} = (-1, 2)^T \text{ and } w_0 = 0$$

Hyperplane



Then, we have

The following results

$$g\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right) = (-1, 2)\begin{pmatrix} 1 \\ 2 \end{pmatrix} = -1 \times 1 + 2 \times 2 = 3$$

$$g\left(\begin{pmatrix} 3 \\ 1 \end{pmatrix}\right) = (-1, 2)\begin{pmatrix} 3 \\ 1 \end{pmatrix} = -1 \times 3 + 2 \times 1 = -1$$

YES!!! We have a positive side and a negative side!!!

Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

The Decision Surface

The equation $g(x) = 0$ defines a decision surface

Separating the elements in classes, ω_1 and ω_2.

When $g(x)$ is linear the decision surface is an hyperplane

Now assume x_1 and x_2 are both on the decision surface

$$w^T x_1 + w_0 = 0$$
$$w^T x_2 + w_0 = 0$$

Thus

$$w^T x_1 + w_0 = w^T x_2 + w_0 \quad (3)$$

The Decision Surface

The equation $g(x) = 0$ defines a decision surface

Separating the elements in classes, ω_1 and ω_2.

When $g(x)$ is linear the decision surface is an hyperplane

Now assume $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both on the decision surface

$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0$$
$$\boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0$$

Thus

$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = \boldsymbol{w}^T \boldsymbol{x}_2 + w_0 \quad (3)$$

The Decision Surface

The equation $g(x) = 0$ defines a decision surface

Separating the elements in classes, ω_1 and ω_2.

When $g(x)$ is linear the decision surface is an hyperplane

Now assume $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both on the decision surface

$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0$$
$$\boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0$$

Thus

$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = \boldsymbol{w}^T \boldsymbol{x}_2 + w_0 \tag{3}$$

Defining a Decision Surface

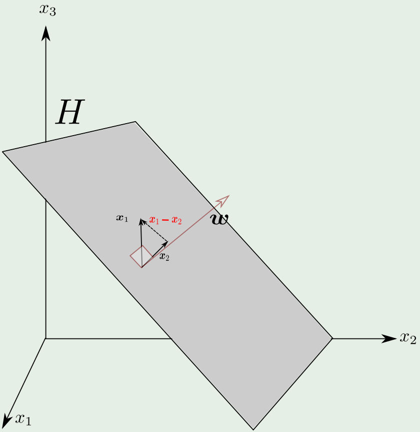
Then

$$\boldsymbol{w}^T \left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right) = 0 \tag{4}$$

Therefore

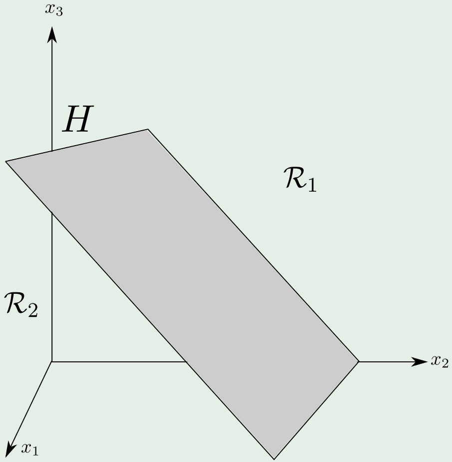
$\boldsymbol{x}_1 - \boldsymbol{x}_2$ lives in the hyperplane i.e. it is perpendicular to $\boldsymbol{w}^T$

- Remark: any vector in the hyperplane is a linear combination of elements in a basis
- **Therefore any vector in the plane is perpendicular to w^T**



Therefore

The space is split in two regions (Example in $\mathbb{R}^3$) by the hyperplane H



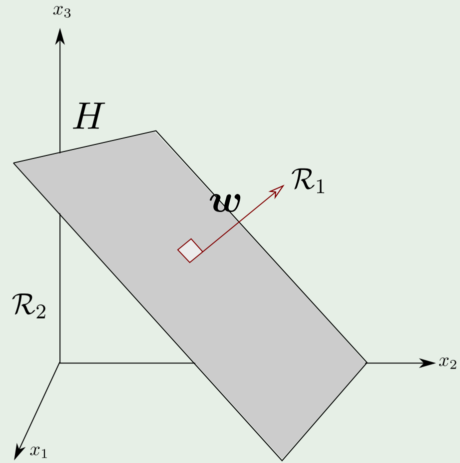
Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $\boldsymbol{w}^T\boldsymbol{x} + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Some Properties of the Hyperplane

Given that $g(\boldsymbol{x}) > 0$ if $\boldsymbol{x} \in \mathcal{R}_1$



It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, $g(x)$ can give us a way to obtain the distance from x to the hyperplane H

First, we express any x as follows

$$x = x_p + r\frac{w}{\|w\|}$$

- r is the desired distance
  - Positive, if x is in the positive side
  - Negative, if x is in the negative side

It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

- x_p is the normal projection of x onto H.
- r is the desired distance
  - Positive, if x is in the positive side
  - Negative, if x is in the negative side

It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if x is in the positive side
  - Negative, if x is in the negative side

It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if x is in the positive side
  - Negative, if x is in the negative side

It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if $\boldsymbol{x}$ is in the positive side
  - Negative, if $\boldsymbol{x}$ is in the negative side

It is more

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

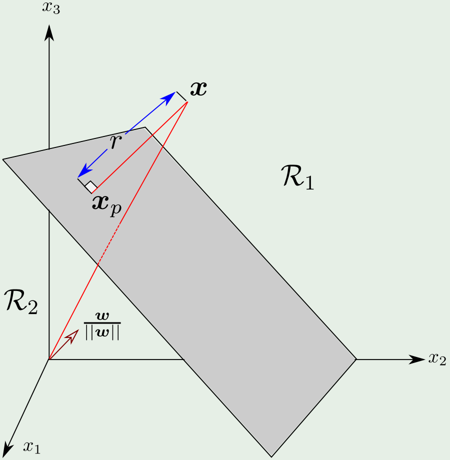
$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if $\boldsymbol{x}$ is in the positive side
  - Negative, if $\boldsymbol{x}$ is in the negative side

We have something like this

We have then



Now

Since $g\left(\boldsymbol{x}_p\right) = 0$

We have that

$$
\begin{aligned}
g\left(\boldsymbol{x}\right) &= g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\left\Vert \boldsymbol{w}\right\Vert}\right)\\
&= \boldsymbol{w}^T\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\left\Vert \boldsymbol{w}\right\Vert}\right) + w_0\\
&= \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\left\Vert \boldsymbol{w}\right\Vert}\\
&= g\left(\boldsymbol{x}_p\right) + r\frac{\left\Vert \boldsymbol{w}\right\Vert^2}{\left\Vert \boldsymbol{w}\right\Vert}\\
&= r\left\Vert \boldsymbol{w}\right\Vert
\end{aligned}
$$

Now

Since $g\left(\boldsymbol{x}_p\right)=0$

We have that

$$\begin{aligned} g\left(\boldsymbol{x}\right) &= g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\left\Vert \boldsymbol{w}\right\Vert}\right)\\ &= \boldsymbol{w}^T\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\left\Vert \boldsymbol{w}\right\Vert}\right) + w_0\\ &= \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\left\Vert \boldsymbol{w}\right\Vert}\\ &= g\left(\boldsymbol{x}_p\right) + r\frac{\left\Vert \boldsymbol{w}\right\Vert^2}{\left\Vert \boldsymbol{w}\right\Vert}\\ &= r\left\Vert \boldsymbol{w}\right\Vert \end{aligned}$$

Then, we have

$$r = \frac{g\left(\boldsymbol{x}\right)}{\left\Vert \boldsymbol{w}\right\Vert} \tag{5}$$

Now

Since $g(\boldsymbol{x}_p) = 0$

We have that

$$
\begin{aligned}
g(\boldsymbol{x}) &= g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \\
&= \boldsymbol{w}^T\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_0 \\
&= \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|} \\
&= g(\boldsymbol{x}_p) + r\frac{\|\boldsymbol{w}\|^2}{\|\boldsymbol{w}\|} \\
&= r\|\boldsymbol{w}\|
\end{aligned}
$$

Then, we have

$$
r = \frac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|} \tag{5}
$$

Now

Since $g\left(\boldsymbol{x}_p\right) = 0$

We have that

$$
\begin{aligned}
g\left(\boldsymbol{x}\right) &= g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)\\
&= \boldsymbol{w}^T\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_0\\
&= \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|}\\
&= g\left(\boldsymbol{x}_p\right) + r\frac{\|\boldsymbol{w}\|^2}{\|\boldsymbol{w}\|}\\
&= r\|\boldsymbol{w}\|
\end{aligned}
$$

Then, we have

$$
r = \frac{g\left(\boldsymbol{x}\right)}{\|\boldsymbol{w}\|} \tag{5}
$$

Now

Since $g\left(\boldsymbol{x}_p\right) = 0$

We have that

$$\begin{aligned} g\left(\boldsymbol{x}\right) &= g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \\ &= \boldsymbol{w}^T\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_0 \\ &= \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|} \\ &= g\left(\boldsymbol{x}_p\right) + r\frac{\|\boldsymbol{w}\|^2}{\|\boldsymbol{w}\|} \\ &= r\left\|\boldsymbol{w}\right\| \end{aligned}$$

Then, we have

$$r = \frac{g\left(\boldsymbol{x}\right)}{\|\boldsymbol{w}\|} \tag{5}$$

Now

Since $g(\boldsymbol{x}_p) = 0$

We have that

$$\begin{aligned} g(\boldsymbol{x}) &= g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \\ &= \boldsymbol{w}^T\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_0 \\ &= \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|} \\ &= g(\boldsymbol{x}_p) + r\frac{\|\boldsymbol{w}\|^2}{\|\boldsymbol{w}\|} \\ &= r\|\boldsymbol{w}\| \end{aligned}$$

Then, we have

$$r = \frac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|} \tag{5}$$

In particular

The distance from the origin to H

$$r = \frac{g(\mathbf{0})}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T(\mathbf{0}) + w_0}{\|\boldsymbol{w}\|} = \frac{w_0}{\|\boldsymbol{w}\|} \tag{6}$$

In particular

The distance from the origin to H

$$r = \frac{g\left(\mathbf{0}\right)}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T\left(\mathbf{0}\right) + w_0}{\|\boldsymbol{w}\|} = \frac{w_0}{\|\boldsymbol{w}\|} \tag{6}$$

Remarks

- If $w_0 > 0$, the origin is on the positive side of H.
- If $w_0 < 0$, the origin is on the negative side of H.
- If $w_0 = 0$, the hyperplane has the homogeneous form $\boldsymbol{w}^T\boldsymbol{x}$ and hyperplane passes through the origin.

In particular

The distance from the origin to H

$$r = \frac{g(\mathbf{0})}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T(\mathbf{0}) + w_0}{\|\boldsymbol{w}\|} = \frac{w_0}{\|\boldsymbol{w}\|} \tag{6}$$

Remarks

- If $w_0 > 0$, the origin is on the positive side of H.
- If $w_0 < 0$, the origin is on the negative side of H.
- If $w_0 = 0$, the hyperplane has the homogeneous form $\boldsymbol{w}^T\boldsymbol{x}$ and hyperplane passes through the origin.

In particular

The distance from the origin to H

$$r = \frac{g(\mathbf{0})}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T(\mathbf{0}) + w_0}{\|\boldsymbol{w}\|} = \frac{w_0}{\|\boldsymbol{w}\|} \tag{6}$$

Remarks

- If $w_0 > 0$, the origin is on the positive side of H.
- If $w_0 < 0$, the origin is on the negative side of H.
- If $w_0 = 0$, the hyperplane has the homogeneous form $\boldsymbol{w}^T\boldsymbol{x}$ and hyperplane passes through the origin.

Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

We want to solve the independence of w_0

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\boldsymbol{x}) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = w_0 \times x_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i \quad (7)$$

We want to solve the independence of w_0

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\boldsymbol{x}) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = w_0 \times x_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i \tag{7}$$

By making

$$x_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

x_{aug} is called an augmented feature vector.

We want to solve the independence of w_0

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\boldsymbol{x}) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = w_0 \times x_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i \qquad (7)$$

By making

$$\boldsymbol{x}_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

Where

$\boldsymbol{x}_{aug}$ is called an augmented feature vector.

We want to solve the independence of w_0

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\boldsymbol{x}) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = w_0 \times x_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i \tag{7}$$

By making

$$\boldsymbol{x}_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \\ \boldsymbol{x} \\ \\ \end{pmatrix}$$

Where

x_{aug} is called an augmented feature vector.

We want to solve the independence of w_0

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\boldsymbol{x}) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = w_0 \times x_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i \tag{7}$$

By making

$$\boldsymbol{x}_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

Where

$\boldsymbol{x}_{aug}$ is called an augmented feature vector.

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} w_0 \\ \\ \boldsymbol{w} \\ \\ \end{pmatrix}$$

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} w_0 \\ \\ \boldsymbol{w} \\ \\ \end{pmatrix}$$

Remarks

- The addition of a constant component to $\boldsymbol{x}$ preserves all the distance relationship between samples.
- The resulting x_{aug} vectors, all lie in a d-dimensional subspace which is the x-space itself.

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} w_0 \\ \\ \boldsymbol{w} \\ \\ \end{pmatrix}$$

Remarks

- The addition of a constant component to $\boldsymbol{x}$ preserves all the distance relationship between samples.
- The resulting $\boldsymbol{x}_{aug}$ vectors, all lie in a d-dimensional subspace which is the $\boldsymbol{x}$-space itself.

Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Initial Supposition

Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2.

Initial Supposition

Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2.

We want a vector weight $\boldsymbol{w}$ such that

- $\boldsymbol{w}^T \boldsymbol{x}_i > 0$, if $\boldsymbol{x}_i \in \omega_1$.
- $\boldsymbol{w}^T \boldsymbol{x}_i < 0$, if $\boldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.

Initial Supposition

Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2.

We want a vector weight $\boldsymbol{w}$ such that

- $\boldsymbol{w}^T \boldsymbol{x}_i > 0$, if $\boldsymbol{x}_i \in \omega_1$.
- $\boldsymbol{w}^T \boldsymbol{x}_i < 0$, if $\boldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.

Initial Supposition

Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2.

We want a vector weight $\boldsymbol{w}$ such that

- $\boldsymbol{w}^T \boldsymbol{x}_i > 0$, if $\boldsymbol{x}_i \in \omega_1$.
- $\boldsymbol{w}^T \boldsymbol{x}_i < 0$, if $\boldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.

Now, assume the following

Imagine that your problem has two classes ω_1 and ω_2 in $\mathbb{R}^2$

1. They are linearly separable!!!
2. You require to label them.

Now, assume the following

Imagine that your problem has two classes ω_1 and ω_2 in $\mathbb{R}^2$

1. They are linearly separable!!!
2. You require to label them.

We have a problem!!!

Which is the problem?

Now, assume the following

Imagine that your problem has two classes ω_1 and ω_2 in $\mathbb{R}^2$

1. They are linearly separable!!!
2. You require to label them.

We have a problem!!!

Which is the problem?

We do not know the hyperplane!!!

Thus, what distance each point has to the hyperplane?

Now, assume the following

Imagine that your problem has two classes ω_1 and ω_2 in $\mathbb{R}^2$

1. They are linearly separable!!!
2. You require to label them.

We have a problem!!!

Which is the problem?

We do not know the hyperplane!!!

Thus, what distance each point has to the hyperplane?

A Simple Solution For Our Quandary

Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

A Simple Solution For Our Quandary

Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

We produce the following labels

1. if $\boldsymbol{x} \in \omega_1$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = +1$.
2. if $\boldsymbol{x} \in \omega_2$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = -1$.

Remark: We have a problem with this labels!!!

A Simple Solution For Our Quandary

Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

We produce the following labels

1. if $\boldsymbol{x} \in \omega_1$ then $y_{ideal} = g_{ideal}\left(\boldsymbol{x}\right) = +1$.
2. if $\boldsymbol{x} \in \omega_2$ then $y_{ideal} = g_{ideal}\left(\boldsymbol{x}\right) = -1$.

Remark: We have a problem with this labels!!!

A Simple Solution For Our Quandary

Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

We produce the following labels

1. if $\boldsymbol{x} \in \omega_1$ then $y_{ideal} = g_{ideal}\left(\boldsymbol{x}\right) = +1$.
2. if $\boldsymbol{x} \in \omega_2$ then $y_{ideal} = g_{ideal}\left(\boldsymbol{x}\right) = -1$.

Remark: We have a problem with this labels!!!

Outline

1 Introduction
- Introduction
- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector

2 Developing a Solution
- Least Squared Error Procedure
  - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation
- Geometric Interpretation
- Issues with Least Squares!!!
  - Problem with Outliers
  - Problem with High Number of Dimensions

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 + e \tag{8}$$

Where the e

It has a $e \sim N(\mu, \sigma^2)$

Thus, we can do the following

$$y_{noise} = g_{noise}(\boldsymbol{x}) = g_{ideal}(\boldsymbol{x}) + e \tag{9}$$

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}\left(\boldsymbol{x}\right) = \boldsymbol{w}^T \boldsymbol{x} + w_0 + e \tag{8}$$

Where the e

It has a $e \sim N\left(\mu, \sigma^2\right)$

Thus, we can do the following

$$y_{noise} = g_{noise}\left(\boldsymbol{x}\right) = g_{ideal}\left(\boldsymbol{x}\right) + e \tag{9}$$

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}\left(\boldsymbol{x}\right) = \boldsymbol{w}^T\boldsymbol{x} + w_0 + e \quad (8)$$

Where the e

It has a $e \sim N\left(\mu, \sigma^2\right)$

Thus, we can do the following

$$y_{noise} = g_{noise}\left(\boldsymbol{x}\right) = g_{ideal}\left(\boldsymbol{x}\right) + e \quad (9)$$

Thus, we have

What to do?

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right) \tag{10}$$

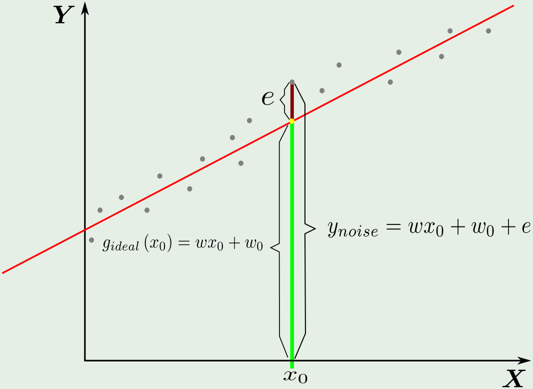
Graphically

Thus, we have

What to do?

$$e = y_{noise} - g_{ideal}(\boldsymbol{x}) \tag{10}$$

Graphically



Then, we have

A TRICK... Quite a good one!!! Instead of using y_{noise}

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right) \tag{11}$$

We use y_{ideal}

$$e = y_{ideal} - g_{ideal}\left(\boldsymbol{x}\right) \tag{12}$$

We will see

How the geometry will solve the problem with using these labels.

Then, we have

A TRICK... Quite a good one!!! Instead of using y_{noise}

$$e = y_{noise} - g_{ideal}(\boldsymbol{x}) \tag{11}$$

We use y_{ideal}

$$e = y_{ideal} - g_{ideal}(\boldsymbol{x}) \tag{12}$$

We will see

How the geometry will solve the problem with using these labels.

Then, we have

A TRICK... Quite a good one!!! Instead of using y_{noise}

$$e = y_{noise} - g_{ideal}(\boldsymbol{x}) \tag{11}$$

We use y_{ideal}

$$e = y_{ideal} - g_{ideal}(\boldsymbol{x}) \tag{12}$$

We will see

How the geometry will solve the problem with using these labels.

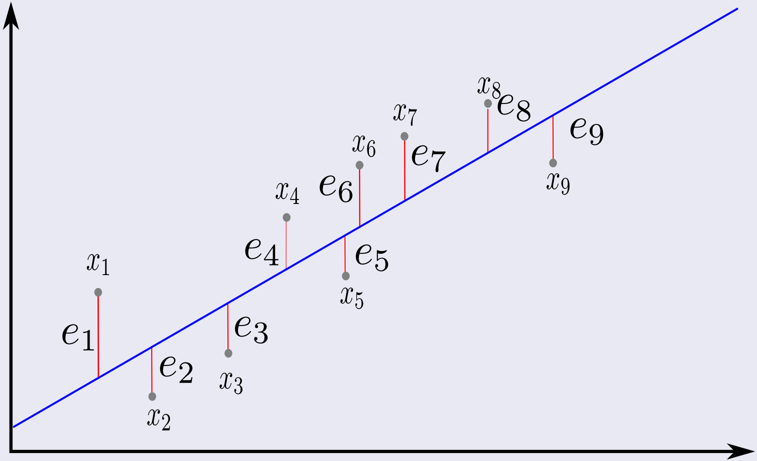
Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Here, we have multiple errors

What can we do?



Sum Over All the Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - g_{ideal}(\boldsymbol{x}_i)\right)^2 \tag{13}$$

Remark: This is know as the Least Squared Error cost function

Sum Over All the Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - g_{ideal}\left(\boldsymbol{x}_i\right)\right)^2 \tag{13}$$

Remark: This is know as the Least Squared Error cost function

Generalizing

- The dimensionality of each sample (data point) is d.
- You can extend each vector sample to be $x^T = (1, x')$.

Sum Over All the Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - g_{ideal}(\boldsymbol{x}_i)\right)^2 \tag{13}$$

Remark: This is know as the Least Squared Error cost function

Generalizing

- The dimensionality of each sample (data point) is d.
- You can extend each vector sample to be $\boldsymbol{x}^T = (\mathbf{1}, \boldsymbol{x}')$.

We can use a trick

The following function

$$g_{ideal}\left(\boldsymbol{x}\right) = \begin{pmatrix} 1 & x_1 & x_2 & \dots & x_d \end{pmatrix} \begin{pmatrix} w_0 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix} = \boldsymbol{x}^T \boldsymbol{w}$$

We can rewrite the error equation as

$$J\left(\boldsymbol{w}\right) = \sum_{i=1}^{N} \left(y_i - g_{ideal}\left(\boldsymbol{x}_i\right)\right)^2 = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\right)^2 \quad (14)$$

We can use a trick

The following function

$$g_{ideal}(\boldsymbol{x}) = \begin{pmatrix} 1 & x_1 & x_2 & \dots & x_d \end{pmatrix} \begin{pmatrix} w_0 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix} = \boldsymbol{x}^T \boldsymbol{w}$$

We can rewrite the error equation as

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} \left(y_i - g_{ideal}(\boldsymbol{x}_i)\right)^2 = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\right)^2 \tag{14}$$

Furthermore

Then stacking all the possible estimations into the product Data Matrix and weight vector

$$\boldsymbol{X}\boldsymbol{w} = \begin{pmatrix} 1 & (\boldsymbol{x}_1)_1 & \cdots & (\boldsymbol{x}_1)_j & \cdots & (\boldsymbol{x}_1)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_i)_1 & & (\boldsymbol{x}_i)_j & & (\boldsymbol{x}_i)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_N)_1 & \cdots & (\boldsymbol{x}_N)_j & \cdots & (\boldsymbol{x}_N)_d \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{d+1} \end{pmatrix}$$

Note about other representations

We could have $\boldsymbol{x}^T = (x_1, x_2, ..., x_d, 1)$ thus

$$\boldsymbol{X} = \begin{pmatrix} (\boldsymbol{x}_1)_1 & \cdots & (\boldsymbol{x}_1)_j & \cdots & (\boldsymbol{x}_1)_d & 1 \\ & & \vdots & & \vdots & \vdots \\ (\boldsymbol{x}_i)_1 & & (\boldsymbol{x}_i)_j & & (\boldsymbol{x}_i)_d & 1 \\ & & \vdots & & \vdots & \vdots \\ (\boldsymbol{x}_N)_1 & \cdots & (\boldsymbol{x}_N)_j & \cdots & (\boldsymbol{x}_N)_d & 1 \end{pmatrix} \tag{15}$$

Then, we have the following trick with $\boldsymbol{X}$

With the Data Matrix

$$\boldsymbol{X}w = \begin{pmatrix} \boldsymbol{x}_1^T\boldsymbol{w} \\ \boldsymbol{x}_2^T\boldsymbol{w} \\ \boldsymbol{x}_3^T\boldsymbol{w} \\ \vdots \\ \boldsymbol{x}_N^T\boldsymbol{w} \end{pmatrix} \tag{16}$$

Therefore

We have that

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_4 \end{pmatrix} - \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{w} \\ \boldsymbol{x}_2^T \boldsymbol{w} \\ \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix}$$

Then, we have the following equality

$$\begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} & y_2 - \boldsymbol{x}_2^T \boldsymbol{w} & y_3 - \boldsymbol{x}_3^T \boldsymbol{w} & \cdots & y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\right)^2$$

Therefore

We have that

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_4 \end{pmatrix} - \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{w} \\ \boldsymbol{x}_2^T \boldsymbol{w} \\ \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix}$$

Then, we have the following equality

$$\begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} & y_2 - \boldsymbol{x}_2^T \boldsymbol{w} & y_3 - \boldsymbol{x}_3^T \boldsymbol{w} & \cdots & y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w} \right)^2$$

Then, we have

The following equality

$$\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{w}\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \quad (17)$$

The Final Discriminant Function

Very Simple!!!

$$g(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w} = \boldsymbol{x}^T \left(\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} \tag{18}$$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\left(X\right) = m$. We call the matrix

$$X^{+} = \left(X^{T}X\right)^{-1}X^{T}$$

the pseudo inverse of X.

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\left(X\right) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\left(X\right) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m\times n}$ and $rank\,(X) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

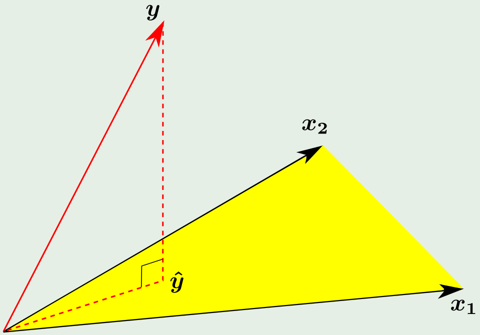
the pseudo inverse of X.

Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector
2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Geometrically

Given a $\boldsymbol{y}$, you obtain a projected $\widehat{\boldsymbol{y}}$ through the process $\boldsymbol{X}^T\boldsymbol{y}$



This Resolve Our Problem

With the Labels being chosen at the beginning

Question? Did you noticed the following?

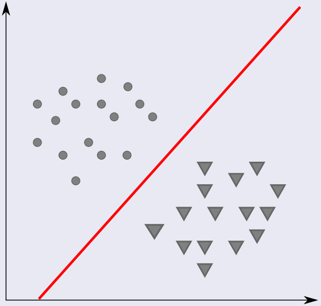
We assume a similar number of elements in both classes

This Resolve Our Problem

With the Labels being chosen at the beginning

Question? Did you noticed the following?

We assume a similar number of elements in both classes



Outline

1 Introduction
- Introduction
- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector

2 Developing a Solution
- Least Squared Error Procedure
  - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation
- Geometric Interpretation
- Issues with Least Squares!!!
  - Problem with Outliers
  - Problem with High Number of Dimensions

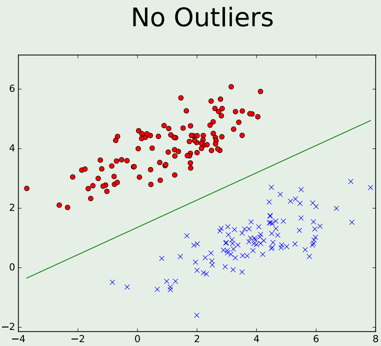
Outline

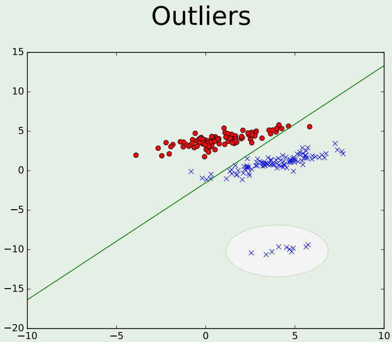
1 Introduction
- Introduction
- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector

2 Developing a Solution
- Least Squared Error Procedure
  - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation
- Geometric Interpretation
- Issues with Least Squares!!!
  - Problem with Outliers
  - Problem with High Number of Dimensions

Issues with Least Squares

Problem with Outliers





Outline

1. Introduction
   - Introduction
   - The Simplest Functions
   - Splitting the Space
   - Defining the Decision Surface
   - Properties of the Hyperplane $w^T x + w_0$
   - Augmenting the Vector

2. Developing a Solution
   - Least Squared Error Procedure
     - The Geometry of a Two-Category Linearly-Separable Case
   - The Error Idea
   - The Final Error Equation
   - Geometric Interpretation
   - Issues with Least Squares!!!
     - Problem with Outliers
     - Problem with High Number of Dimensions

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

- Possible some type of regularization

- Relevant dimensions might be averaged with irrelevant ones.

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

- Relevant dimensions might be averaged with irrelevant ones.

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.